DOI:10.11721/cqnuj20230514

层次预处理的非负矩阵分解加权集成聚类算法*

李向利[1,2], 毕 胜[1,3], 王佩源[1]
(1. 桂林电子科技大学 数学与计算科学学院; 2. 广西高校数据分析与计算重点实验室;
3. 广西应用数学中心, 广西 桂林 541004)

摘要:图像聚类是当前的研究热点,非负矩阵分解(non-negative matrix factorization, NMF)算法在图像聚类领域得到了广泛应用。但是单一的 NMF 算法无法应用于所有数据集,并且 NMF 算法直接在数据的原始空间进行处理,抗噪能力较差。集成聚类可以解决上述问题,集成聚类将若干个基础聚类结果合成一个一致性结果,不仅可以提高聚类的求解质量,还可以增强算法的鲁棒性。因此本文提出一种层次预处理的 NMF 加权集成聚类算法。该算法将层次划分、集成聚类和二部图的思想引入到 NMF 算法中。在预处理阶段,利用层次划分得到聚类数目。之后采用局部加权的方法得到协关联矩阵。最后利用基于二部图的一致性函数进行划分得到最终的聚类结果。在 5 个数据集上进行实验,验证了本文算法相对于传统算法和其他集成算法的有效性。

关键词:图像聚类;聚类集成;非负矩阵分解
中图分类号:O151.21;TP391.4 **文献标志码**:A **文章编号**:1672-6693(2023)05-0136-09

随着信息技术的飞速发展,数据形式和种类呈现出多样化特点。如何更好地对海量数据进行管理并挖掘成为科学研究的核心内容之一。数据聚类一直是机器学习、模式识别及数据挖掘等领域的重要组成部分,并逐渐成为机器学习中应用最广泛的技术之一。数据聚类的目标是发现数据集中的自然分组[1],可以定义为将对象分到一些簇中,使簇内的对象有很高的相似度,簇间的对象相似度则较低。数据聚类通常被用于发现数据的潜在结构、对数据进行自然的分组以及对数据进行压缩等方面。常用的聚类方法有:基于划分的聚类,如 K-means 算法[2];基于密度的聚类算法,如 DBSCAN 算法[3];基于网格的聚类算法,如 GRIDCLUS 算法[4];基于层次的聚类算法,如 SLINK 算法[5]等。图像数据的规模和维数在大数据技术和自媒体的发展下日趋增加,因此图像聚类成为新的研究热点。常用降维分解的方法对它进行研究,目前已经提出了很多种有效的降维聚类算法,如:主成分分析(principal component analysis, PCA)[6]、奇异值分解(singular value decomposition, SVD)[7]等。由于图像数据通常以非负的方式储存,因此使用 Lee 等人[8]提出非负矩阵分解(non-negative matrix factorization, NMF)算法更加符合现实情况,因为经过分解后得到的矩阵依然为非负,这增强了图片数据集转换为矩阵后的可解释性。

然而,单一的聚类算法往往存在较大的局限性,不同的聚类算法对相同数据集进行聚类可能会产生较大的差异性,没有任何一种单一聚类算法可以适用于所有的数据集。集成聚类的出现很好地解决了此问题,它将不同的聚类算法结果或同一算法的多次聚类结果进行合并,最终形成一个一致性的聚类结果的算法,且该一致性结果均好于集成中的单一聚类算法。集成聚类技术由 Strehl 等人[9]提出,并证明了集成聚类的基本聚类结果具有互补性。与单一聚类算法相比,集成聚类可以明显提高聚类的性能、鲁棒性和稳定性。

近些年来,关于集成聚类的研究工作取得了一些成果。例如:U-SENC[10]、MDEC[11]等方法使高维大规模数据可以在单机环境下进行集成聚类,此外还有基于模糊聚类的集成算法[12]、基于谱聚类的集成算法[13]等。但是,目前仍然存在一些亟待解决的问题:1) 在聚类实验中,需要规定聚类数目,实验中通常需要使用真实的聚类数目,但实际应用中却很难得到这类数据;2) 在集成聚类中,不同的基本聚类所产生的结果并不相同,对最终的

* **收稿日期**:2022-04-11 **修回日期**:2022-12-14 **网络出版时间**:2023-09-22T11:31
资助项目:国家自然科学基金面上项目(No. 11961010;No. 61967004)
第一作者简介:李向利,女,教授,博士,研究方向为最优化算法、图像聚类、非负矩阵分解等,E-mail: lixiangli@guet.edu.cn
网络出版地址:https://link.cnki.net/urlid/50.1165.N.20230921.1407.002

一致性聚类结果造成的影响也不同，如何平衡不同基本聚类对最终结果的影响是提升一致性聚类结果的关键。

为解决上述问题，受非负矩阵分解和集成聚类算法的启发，本文提出了一种层次预处理的 NMF 加权集成聚类，分析数据集的层次划分，并且使用熵值计算可靠性指数[14]进行局部加权。本文的主要贡献包括：1）利用数据的层次特性在聚类前对数据进行层次划分，可以代替在实验中直接使用真实类别数，增强了仿真实验应用价值；2）将不同类型的 NMF 算法进行集成，并通过对不同方法加权使聚类结果在保持原有聚类特性的基础上又能消除对不适用簇的影响；3）利用基于二部图的一致性函数，将数据的样本点和基簇作为图节点建立二部图，然后进行二部图的划分，得到最终的一致性聚类结果，明显降低了算法的时间成本。

1　相关研究

目前关于 NMF 算法聚类的研究各有不相同的优势和特点。如正交非负矩阵分解（orthogonal non-negative matrix factorization，ONMF）被证明与 K-means 算法等价[15]；正交对称非负矩阵分解（orthogonal symmetric non-negative matrix factorization，OSNMF）被证明与谱聚类等价[16]；投影非负矩阵分解（projective NMF）方法[17]，试图找到一个投影矩阵，把原始数据映射到一个低维空间，它的思想和子空间聚类相似，时间复杂度也较低，但由于投影非负矩阵分解方法在非凸理论方面欠缺，因此该算法缺少极值收敛性分析；核非负矩阵分解（Kernel-NMF）方法[18]的目标公式不再与核函数 $\varphi(\cdot)$ 相关，而是与核矩阵 $\varphi(\cdot)^{\mathrm{T}}\varphi(\cdot)$ 相关，更容易求解分解后的子矩阵，但是 Kernel-NMF 方法没有对特征矩阵和基本图像进行约束，存在鲁棒性较差的问题；此外还有基于局部约束自适应图的非负矩阵分解（NMF-LCAG）[19]，该方法能够发现数据潜在的流形结构，但是实验结果波动较大，容易陷入局部最优，需要大量的实验才能得到趋于稳定的结果；双流形正则化的非负矩阵分解（DMR-NMF）[20]方法中的局部几何结构可以使学习到的表示更有效。

近些年来集成聚类受到了研究者的广泛关注，因为它可提高聚类的求解质量和鲁棒性。集成聚类使用基本聚类得到一个一致性聚类结果，它不需要访问原始数据，因此集成聚类具有保护隐私的特点。此外，它在并行计算、数据重用方面有着很强的应用价值。到目前为止，现有的集成聚类大致可以分为：1）基于图的集成聚类：如 GUO 等人[21]利用相似图进行集成，充分利用了基本聚类结果之间的多尺度间接关系；Liu 等人[22]利用谱集成聚类，并在理论上证明了该方法与加权均值算法的等价性。2）基于相似矩阵的集成聚类，如 Huang 等人[23]利用基于转移概率矩阵的随机游走过程对集成过程进行更新。3）基于概率统计的集成聚类，如 Wang 等人[24]将贝叶斯理论引入集成聚类框架；Topchy[25]将互信息引入到集成中，核心思想是在基本聚类中找到互信息最大的聚类。此外还有一些集成聚类方法与上述 3 种方法不同，如 Yang 等人[26]将集成聚类视为一个优化问题，并将遗传算法的思想融入集成聚类，遗传算法可以使聚类的一致性结果避免陷入局部最优。

集成聚类算法研究是为了进一步提高聚类结果，它的优势之一是可以利用基本聚类的特点得到一个一致性结果，但是由于不同的基本聚类算法对于数据的适应性并不相同，如果过分考虑基本聚类结果而不作进一步处理，可能会使最终的一致性结果低于任何一次基本聚类结果。因此如何将基本聚类有效的集成是集成聚类的关键。

2　提出算法

本节将详细说明所提出的算法，具体来说，先对目标数据集进行层次分析，得到实验环节所需要的类别数，然后将 NMF 算法进行集成，对基本聚类进行加权，最后构造二部图，利用转移切割的方法得到最终的一致性聚类结果。

2.1　层次聚类

通过层次聚类[27]（hierarchical clustering）的思想对算法输入的数据进行处理，有别于一般的层次聚类[27]，本文对数据进行层次聚类的目的是为了得到输入数据的层次划分而不是聚类结果，因此不设定阈值。数据点间的距离通常采用欧式距离作为度量，也可以根据所设计的具体问题设置其他的度量方式。层次聚类初始时是将每个样本点视为一个簇，然后计算簇间距离，并将距离最小的 2 个簇合并为 1 个新簇，直到将整个数据集凝聚为 1 个大簇。由于层次聚类一次性得到了整个聚类过程，因此将层次聚类过程中每次迭代后各个簇间的最小距离

进行统计绘制,可以得到明显的层次划分,以划分结果作为数据集的聚类数目。例如图1的玩具数据集,对该数据集进行层次聚类,随机选取一些样本点绘制自下而上的树状图(图2),可以清晰看到数据集的层次结构,将得到的层次数量记为 k。

设算法输入数据为 D,其中的样本量为 n,样本用 $\{c_1,\cdots,c_n\}$ 表示,层次聚类的流程如下[28]:

1. 将样本点视为 n 个独立的簇;

repeat:

2. 计算两两簇之间的距离,可以找到距离最小的2类簇 c_1 和 c_2;

3. 合并簇 c_1 和 c_2 为1个簇;

until:达到设定的阈值或全部样本被凝聚为1个簇。

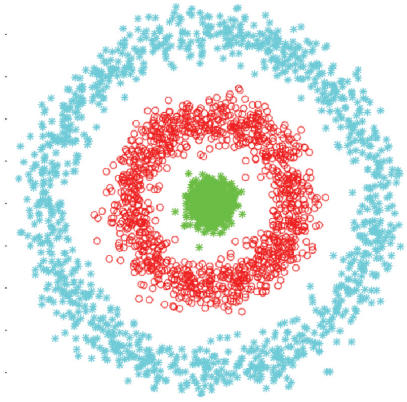

图1 层次划分明显的玩具数据集

Fig. 1 The toy data set which shows a very clear hierarchical division

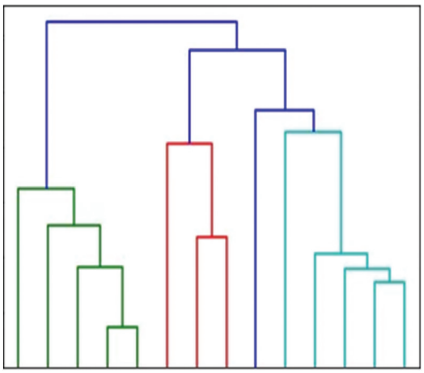

图2 玩具数据集的层次结构树状图

Fig. 2 Hierarchical tree of toy dataset

2.2 基于NMF算法的聚类集成

NMF是一种有效的聚类方法。在采用欧式距离为误差衡量标准时,所求问题的模型[8]为:

$$\min \|\boldsymbol{X}-\boldsymbol{W}\boldsymbol{H}^{\mathrm{T}}\|_F^2,$$

$$\text{s. t. } \boldsymbol{W}\geqslant 0,\boldsymbol{H}\geqslant 0。$$

从聚类的角度考虑,$\boldsymbol{X}\geqslant 0$,$\boldsymbol{X}\in\mathbf{R}^{m\times n}$ 表示以非负矩阵形式存储的数据集;$\boldsymbol{W}\geqslant 0$,$\boldsymbol{W}\in\mathbf{R}^{m\times k}$ 可以理解为所有聚类中心构成的矩阵,一般称为特征矩阵;而 $\boldsymbol{H}\geqslant 0$,$\boldsymbol{H}\in\mathbf{R}^{n\times k}$ 为数据的划分,一般称为系数矩阵。k 表示数据所含的类别数即层次划分数量,本文中类别数均为层次划分得到。

将 M 次不同的NMF方法作为基本聚类,第 m 次基本聚类记为 $\pi^m=\{C_1^m,C_2^m,\cdots,C_k^m\}$,其中:$C_i^m$ 表示在 π^m 中的第 i 个簇,本文中同一个数据在一次基本聚类中只会被分配到1个簇中,即同一个基本聚类中2个簇之间不会重叠 $\forall i\neq j$,$C_i^m\cap C_j^m=\varnothing$。

于是得到1个基本聚类的集成,表示为 $\Pi=\{\pi^1,\pi^2,\cdots,\pi^m\}$,其中:$\pi^m$ 表示为集成中的第 m 次基本聚类。

2.3 构造共识函数

本节的目标是在一个统一的集成框架中充分利用基本聚类的多样性得到一个效果好于基本聚类结果的最终共识聚类结果。

集成中的每个基本聚类都由一定数量的簇组成,因此,可将集成整体看作一个更大的簇的集合。为了利用不同基本聚类的特点,分析不同基本聚类的不同可靠性,引入熵的概念,并采用局部加权策略,对基本聚类进行评估和加权。

首先,将集成中基本簇的集合表示为 $C=\{C_1,C_2,\cdots,C_K\}$,C_i 表示第 i 个簇,K 是集成中簇的总数,记为 $K=\sum_{m=1}^{M}k$。熵代表了数据的不确定性,值越小则结果越可靠。利用熵值的这个性质来对集成中的簇进行加权。对于一个基本聚类 $\pi^m\in\Pi$ 和给定的簇 $C_i\in C$。熵的计算公式为:

$$H^{\pi^m}(C_i)=-\sum_{j=1}^{k}p(C_i,C_j^m)\log_2 p(C_i,C_j^m),$$

其中:$p(C_i,C_j^m)=\dfrac{|C_i\cap C_j^m|}{|C_i|}$表示 C_i 中的数据在 C_j^m 中出现的比例。显然 $p(C_i,C_j^m)\in[0,1]$,从而得到 $H^{\pi^m}(C_i)\in[0,+\infty)$。在基本聚类相互独立的前提下可以得到整个集成的熵值:

$$H^{\Pi}(C_i)=\sum_{m=1}^{M}H^{\pi^m}(C_i)。$$

接下来对熵值进行分析。如果得到较高的值则意味着集群中的基本聚类结果与簇所代表的结果不一致,因此赋予该基本聚类一个较小的权值。进一步利用熵值来计算可靠性指数[29] $\mathrm{ECI}(C_i)=\exp\left(-\dfrac{H^{\Pi}(C_i)}{M}\right)$,并利用它作为共识函数的权重项。显然,对于任意的 $\pi^m\in\Pi$,都可以得到 $H^{\pi^m}(C_i)\in[0,+\infty)$,因此可以得到 $H^{\Pi}(C_i)\in[0,+\infty)$和 $\mathrm{ECI}(C_i)\in[0,1]$。

ECI 的值与熵所代表的不确定性相关,值越大,则不确定性越低,即可靠性较高。因此将 ECI 作为集成中不同簇的可靠性指标,ECI 作为聚类的加权项,得到局部加权的协关联矩阵:

$$\boldsymbol{A}=\{A_{ij}\}_{n\times n},\ A_{ij}=\frac{1}{M}\cdot\sum_{m=1}^{M}w_i^m\cdot\delta_{ij}^m,$$

其中:$w_i^m=\mathrm{ECI}(\mathrm{Cls}^m(x_i))$表示权重项,它根据 ECI 的值对每个簇进行加权;$\delta_{ij}^m=\begin{cases}1,\mathrm{Cls}^m(x_i)=\mathrm{Cls}^m(x_j)\\0,\text{否则}\end{cases}$为共现项,表示 2 个样本点在一个基本聚类 π^m 中是否属于同一个簇,其中$\mathrm{Cls}^m(x_i)$表示数据 x_i 在 π^m 中所属的簇。

在得到局部加权的协关联矩阵后,利用基于二部图的一致性函数,将数据的样本点和基簇作为图节点建立二部图,然后进行二部图的划分,得到最终的一致性聚类结果。具体来说,二部图定义为 $G=(U,V,E)$,$U=X$,$V=C$。在二部图中 $U\cup V$ 是节点集,E 为边集,当一个节点为数据样本,而另一个节点是包含该数据样本的基本簇时,2 个节点之间存在边。二部图的结构如图 3 所示。

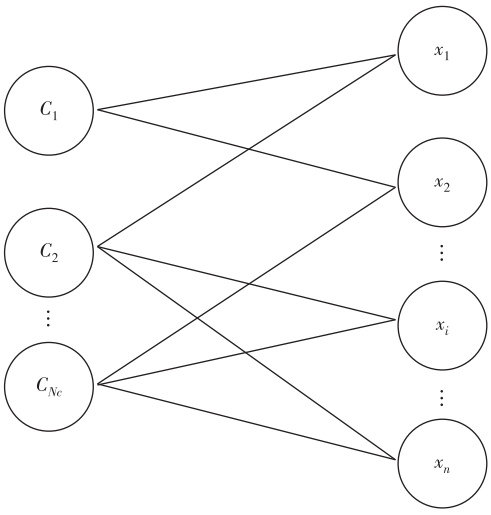

图 3　二部图的一般结构

Fig. 3　General structure of bipartite graphs

通常情况下,给定 2 个节点 $u_i\in X$,$v_j\in C$,它们的边权重由归属关系和所连接的簇的可靠程度决定。这 2 个因素均可以用 ECI 指数来估算,因此将 u_i、v_j 的边权重定义为 $E_{ij}=\begin{cases}\mathrm{ECI}(v_j),u_i\in v_j\\0,\text{否则}\end{cases}$,ECI 反映集成中一个基本聚类的簇的可靠性。因此可靠性较高的簇能对二部图产生更大的影响,从而在形成共识的过程中可以发挥更大的作用,最后在构造二部图的基础上利用转移切割的方法[30]可以有效地将图节点划分为多个子集,也即是一致性聚类最终结果,每个子集中的样本就是最终的聚类。

3 数值实验

3.1 实验数据

数值实验在5个数据集上进行,分别是COIL20数据集[31]、YaleB32数据集[32]、USPS数据集[33]、Flowers17数据集[34]和MNIST数据集[35]。COIL20数据集包含20个物体对象,每个对象72幅图,共360张图片,图片大小为64×64像素。YaleB32数据集包含38个类别,每个类别59幅图片,图片大小为32×32像素。USPS数据集包含10个数字手写体,每个类别200张,图片大小为18×18像素。Flowers17数据集包含17种花,每种花有80幅图片,图片尺寸不一。MNIST数据集包含10个类别,5 000幅图片,图片大小为28×28像素。

实验中,对照选取了单一聚类的NMF算法、K-means算法、谱聚类算法和集成的self-paced clustering ensemble(SPCE)算法[36]、K-means based consensus clustering(KCC)算法[37]、NMF Based K-means clustering ensemble(NBKCE)算法[38]和selective clustering ensemble(SCE)算法[39]。此外,深度学习和聚类相结合也产生了一个新的研究方向,即深度聚类(deep clustering)。本文对比了基于深度聚类的模因群聚类和深度信念网络(memetic swarm clustering and deep belief network, MSC-DBN)[40]方法,该方法主要由聚类、分类和推荐共3个主要阶段构成,由于深度聚类方法与传统聚类方法的结构有所不同,因此在MSC-DBN实验的第一阶段使用蚁群、粒子群优化等方法进行聚类时,参数设置与所参考的文献一致,再进行聚类准确率的对比,以避免参数选取对最终聚类结果所产生的影响。此外,对于深度聚类方法所拥有的上下文感知(CA)、协作过滤(CF)等模块化模型由于结构不同无法与本文中其他方法进行对比。

3.2 评价标准

在衡量聚类结果时采用归一化互信息(normalized mutual information, NMI)[41]和调兰德指数(adjusted rand score, ARI)[42]作为评价指标,分别记作 x_{NMI} 和 x_{ARI}。NMI和ARI的值越大,聚类结果越好。

归一化互信息是衡量聚类结果的标准,计算方法为 $x_{\mathrm{NMI}}=\dfrac{\sum_{i=1,j=1}^{k} n_{i,j}\log\dfrac{n_{i,j}}{n_i\hat{n}_j}}{\sqrt{\left(\sum_{i=1}^{k} n_i\log\dfrac{n_i}{n}\right)\left(\sum_{j=1}^{k} n_j\log\dfrac{\hat{n}_j}{n}\right)}}$,其中:$k$ 表示类别个数,n_i 表示属于第 C_i 个聚类类别的样本数,$\hat{n}_j$ 表示属于真实标签 L_j 的数据个数,$n_{i,j}$ 表示同时属于 C_i 与 L_j 的样本数目。

调兰德指数反映2种划分的重叠程度,计算方法为 $x_{\mathrm{ARI}}=\dfrac{x_{\mathrm{RI}}-E(x_{\mathrm{RI}})}{\max(x_{\mathrm{RI}})-E(x_{\mathrm{RI}})}$,其中,$x_{\mathrm{RI}}$ 为兰德系数(rand index,RI),取范围为[0,1],值越大意为着聚类结果与真实情况越吻合。

3.3 实验对比

实验中各方法参数均与原文献所给参数一致,类别数为对数据集进行层次划分所得类别数。提出算法所使用的NMF算法为ONMF、Projective NMF和Kernel-NMF在随机初始值下作为基本聚类所集成的聚类结果。为避免随机性的影响,对各算法分别进行10次实验再取平均值。各算法的聚类结果的NMI和ARI均值如表1和表2所示。

在NMI方面,实验结果如表1所描述,本文算法在大多数数据集上的结果优于基本算法和对比的集成算法。在Flowers17数据集上本文算法效果不理想,原因可能是未能充分考虑到一些困难实例的不利影响,无法对图片大小不一致的数据进行处理,而MSC-DBN算法在此数据集表现优秀,可能是因为它使用DBN的学习能力对Flowers17数据集进行了更深的分析,这也将是本文应该进一步加强的研究方向。KCC算法和NBKCE算法在某些数据集上表现出性能相似的聚类结果,可能是由于两者都是基于 K-means算法所提出的改进算法。

在ARI方面,由于ARI的标准是划分的重叠程度,因此在某些结果上呈现出与NMI不同的实验结果,但是本文算法依然展现出优于大多数实验方法的结果。KCC算法在YaleB32数据集上有突出的表现,可能是由于KCC算法中所强调的一致性(consensus)在该数据集上有更好的适应性;SCE方法在COIL20数据集上的表现非常优秀,可能是因为SCE算法中基于Kappa和F-Score的选择性方法在COIL20数据中能发挥更大的作用。

表 1　各算法 10 次实验的 NMI 平均结果比较

Tab. 1　Comparison of NMI average results for 10 experiments of each algorithm

	K-means	NMF	Spectral	NBKCE	KCC	SPCE	SCE	MSC-DBN	本文算法
COIL20	53.35	55.87	56.73	77.33	77.49	87.73	89.23	90.35	**90.75**
YaleB32	14.48	15.78	17.23	26.01	34.93	31.99	33.75	34.23	**36.94**
USPS	55.22	58.54	60.59	63.86	65.11	—	73.21	70.53	**74.77**
Flowers17	17.29	18.39	18.32	25.06	24.87	27.17	27.03	**27.29**	26.94
MNIST	41.42	43.36	44.84	60.13	59.06	67.13	75.39	77.49	**77.92**

注：加黑的数据为最优结果，“—”表示数据集在该算法上不可行，下同。

表 2　各算法 10 次实验的 ARI 平均结果比较

Tab. 2　Comparison of ARI average results for 10 experiments of each algorithm

	K-means	NMF	Spectral	NBKCE	KCC	SPCE	SCE	MSC-DBN	本文算法
COIL20	48.85	50.67	53.23	60.73	52.53	64.99	**89.23**	80.05	80.21
YaleB32	9.43	9.68	9.84	13.01	**15.61**	14.41	11.53	12.06	14.40
USPS	17.48	21.25	20.64	34.65	33.41	—	58.25	50.28	**61.88**
Flowers17	6.88	7.98	8.53	9.22	9.51	**10.17**	9.93	9.85	10.08
MNIST	25.63	34.57	32.38	40.53	40.52	34.87	59.07	66.95	**68.42**

3.4　消融实验

为了进一步验证本文算法各个部分对算法性能的贡献，进行了 2 个对比实验。

1）为了验证 ECI 对于实验结果的影响，将集成阶段的基本聚类权重均设为 1，与算法在不加权情况下的聚类指标 NMI 的值比较。对比结果如图 4 所示。可以看出，本文提出的加权集成算法聚类效果均好于不加权情况下的聚类效果，这说明提出的算法可以使聚类结果在保持原有聚类特性的基础上消除对不适用簇的影响，从而得到一个更好的聚类结果。

2）为了验证使用二部图作为一致性函数的作用，设置一种算法在聚类划分阶段不使用二部图，而是进行传统的 K-means 算法划分，对比算法在 2 种情况下的时间，结果如图 5 所示。可以看出，使用二部图作为一致性函数时，实验时间均短于不使用二部图作为一致函数的方法，尤其是在 Flowers17 数据集和 MNIST 数据集上效果突出，原因可能是由于 NMIST 等数据集规模较大，常规的一致性函数生成阶段会导致指数爆炸，最终使聚类时间大大增加。

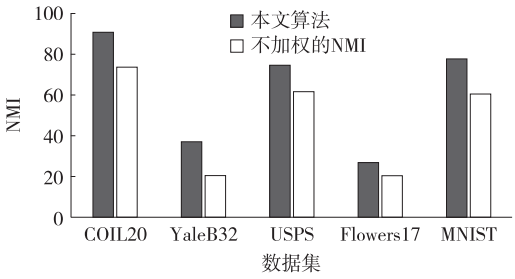

图 4　本文算法与不加权的 NMI 对比

Fig. 4　Propose an algorithm to compare with NMI without weighting

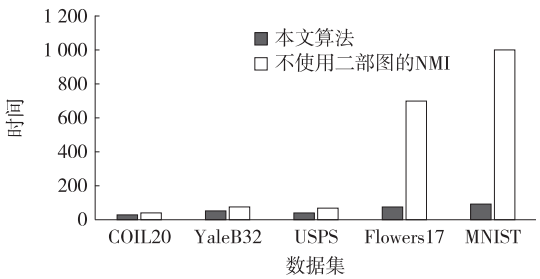

图 5　本文算法与不使用二部图作为一致函数情况下的时间对比

Fig. 5　Comparing the time of the proposed algorithm with that of the non-binary graph as a consistent function

4　结束语

针对图像聚类问题，本文提出了一种层次预处理的 NMF 加权集成聚类算法，该方法利用层次划分对图像数

据集类别数目敏感的特性,可以得到图像数据集的聚类数目。采用NMF聚类集成的思想是为了更好的利用不同NMF特性,同时利用熵的思想进行加权,消除了算法在低效果数据集上产生的负面影响,得到局部加权的协关联矩阵。最后利用一种基于二部图的一致性函数,将数据的样本点和基簇作为图节点建立二部图,采用归一化割的方法进行切割,得到最终的聚类结果。在5个数据集上将5种算法进行了对比,实验说明了本文算法的有效性。但是在实验数据图片大小不一致时,该算法的表现差强人意。此外,集成聚类的复杂度较高,在进行集成时容易出现内存不足的情况,这可能会影响集成聚类的实用性。因此下一步需要对此进行深入研究。

参考文献:

[1] JAIN A K. Data clustering:50 years beyond *K*-means[J]. Pattern Recognition Letters,2010,31(8):651-666.

[2] Mac QUEEN J. Some methods for classification and analysis of multivariate observations[C]//Le CAM L M, NEYMAN J. Proceeding of the 5th of Berkeley Symposium on Mathematical Statistics & Probability. Berkeley:Universitv of California Press, 1965.

[3] ESTER M,KRIEGEL H-P,SANDER J,et al. A density-based algorithm for discovering clusters in large spatial databases with noise[C]//KAUFMAN K A,MICHALSKI R S. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. Menlo Park:AAAI Press 1996.

[4] STEPHEN C J. Hierarchical clustering schemes[J]. Psychometrika,1967,32(3):241-254.

[5] SIBSON R. SLINK:an optimally efficient algorithm for the single-link cluster method[J]. The Computer Journal,1973,16(1):30-34.

[6] BELHUMEUR P N, HESPANHA J P, KRIEGMAN D J. Eigenfaces vs. Fisherfaces:recognition using class specific linear projection[C]//BUZTON B,CIPOLLA R. Computer Vision-ECCV '96. New York:Springer-Verlag,2006.

[7] DAN K. A singularly valuable decomposition:the SVD of a matrix[J]. College Mathematics Journal,1996,27(1):2-23.

[8] LEE D D,SEUNG H S. Learning the parts of objects by non-negative matrix factorization[J]. Nature,1999,401(6755):788-791.

[9] STREHL A, GHOSH J. Cluster ensembles-a knowledge reuse framework for combining multiple partitions[J]. Journal of Machine Learning Research,2002,3(3):583-617.

[10] HUANG D,WANG C D,WU J S,et al. Ultra-scalable spectral clustering and ensemble clustering[J]. IEEE Transactions on Knowledge & Data Engineering,2020,32(6):1212-1226.

[11] HUANG D,WANG C D,LAI J H, et al. Toward multi-diversified ensemble clustering of high-dimensional data: from subspaces to metrics and beyond[J]. IEEE Transactions on Cybernetics,2022,52(11):12231-12244.

[12] STREHL A,GHOSH J. Cluster ensembles:a knowledge reuse framework for combining multiple partitions[J]. Journal of Machine Learning Research,2003,3(3):583-617.

[13] ZHANG X R,JIAO L C,LIU F,et al. Spectral clustering ensemble applied to SAR image segmentation[J]. IEEE Transactions on Geoscience and Remote Sensing,2008,46(7):2126-2136.

[14] CHARNIAK E. A maximum-entropy-inspired parser[D]. Providence:Brown University,1999.

[15] DING C,HE X F,SIMON H D,et al. On the equivalence of nonnegative matrix factorization and spectral clustering[C]//KARGUPTA H, SRIVASTAVA J, KAMATH C, et al. Proceeding of the 2005 SIAM International Conference on Data Mining. Philadelphia:SIAM,2005.

[16] KUANG D,DING C,PARK H. Symmetric nonnegative matrix factorization for graph clustering[C]//GHOSH J,LIU H,DAVIDSON I,et al. Proceeding of the 2012 SIAM International Conference on Data Mining. Philadelphia:SIAM,2012.

[17] YANG Z R, OJA E. Linear and nonlinear projective nonnegative matrix factorization[J]. IEEE Transactions on Neural Networks,2010,21(5):734-749.

[18] RAJKUMAR R, ALBERT A J, CHANDRAKALA D. A new approach to adaptive neuro-fuzzy modeling using kernel nonnegative matrix factorization (KNMF) clustering for weather forecasting[J]. Journal of Advanced Research in Dynamical and Control Systems,2017,9(14):798-826.

[19] DING C H Q,LI T,JORDAN M I. Convex and semi-nonnegative matrix factorizations[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2010,32(1):45-55.

[20] HAN Y F，MOUTARDE F. Analysis of network-level traffic states using locality preservative non-negative matrix factorization [C]//2011 14th International IEEE Conference on Intelligent Transportation Systems，Nov 17，2011，Washington. Piscataway：IEEE，2011.

[21] GUO J P，YIN S，SUN Y F，et al. Double manifolds regularized non-negative matrix factorization for data representation[C]//2020 25th International Conference on Pattern Recognition，January 10-15，2021，Milan，Italy. Piscataway：IEEE，2021.

[22] LIU H F，WU J J，LIU T L，et al. Spectral ensemble clustering via weighted K-means：theoretical and practical evidence[J]. IEEE Transactions on Knowledge and Data Engineering，2017，29(5)：1129-1143.

[23] HUANG D，WANG C D，PENG H X，et al. Enhanced ensemble clustering via fast propagation of cluster-wise similarities[J]. IEEE Transactions on Systems，Man，and Cybernetics：Systems，2021，51(1)：508-520.

[24] WANG H J，SHAN H H，BANERJEE A. Bayesian cluster ensembles[J]. Statistical Analysis and Data Mining，2011，4(1)：54-70.

[25] TOPCHY A，JAIN A K，PUNCH W. Clustering ensembles：models of consensus and weak partitions[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2005，27(12)：1866-1881.

[26] YANG W L，ZHANG Y H，WANG H J，et al. Hybrid genetic model for clustering ensemble[J]. Knowledge-Based Systems，2021，231(4)：107457.

[27] KOGA H，ISHIBASHI T，WATANABE T. Fast agglomerative hierarchical clustering algorithm using Locality-Sensitive Hashing[J]. Knowledge & Information Systems，2007，12(1)：25-53.

[28] BOECKER A，DERKSEN S，SCHMIDT E，et al. A hierarchical clustering approach for large compound libraries[J]. Journal of Chemical Information and Modeling，2005，45(4)：807-815.

[29] DEFAYS D. An efficient algorithm for a complete link method[J]. The Computer Journal，1977，20(4)：364-366.

[30] von LUXBURG U. A Tutorial on spectral clustering[J]. Statistics and Computing，2004，17(4)：395-416.

[31] RATE C，RETRIEVAL C. Columbia object image library (COIL-20)[DB/OL]. [2022-04-09]https：//www. cs. columbia. edu/CAVE/software/softlib/coil-20. php.

[32] 朱凤梅，张道强. 一种基于半监督降维的聚类算法[J]. 广西师范大学学报(自然科学版)，2008，26(3)：185-188.
ZHU F M，ZHANG D Q. A clustering algorithm based on semi-supervised dimensionality reduction[J]. Journal of Guangxi Normal University (Natural Science Edition)，2008，26(3)：185-188.

[33] CHOW W，OGG C. Secured centralized public key infrastructure：US20020178354 A1[P]. 2002.

[34] NILSBACK M E，ZISSERMAN A. 17 Category Flower Dataset[EB/OL]. [2022-04-09]. https：//www. robots. ox. ac. uk/～vgg/data/flowers/17/.

[35] XIAO H，RASUL K，VOLLGRAF R. Fashion-MNIST：a novel image dataset for benchmarking machine learning algorithms [EB/OL]. (2017-09-15)[2022-04-09]. https：//arxiv. org/pdf/1708. 07747. pdf.

[36] ZHOU P，DU L，LIU X W，et al. Self-paced clustering ensemble[J]. IEEE Transactions on Neural Networks and Learning Systems，2020，32(4)：1497-1511.

[37] WU J J，LIU H F，XIONG H，et al. K-means-based consensus clustering：a unified view[J]. IEEE Transactions on Knowledge and Data Engineering，2015，27(1)：155-169.

[38] 何梦娇，杨燕，王淑营. 一种基于非负矩阵分解的聚类集成算法[J]. 计算机科学，2017，44(9)：58-61.
HE M J，YANG Y，WANG S Y. NMF-based clustering ensemble algorithm[J]. Computer Science，2017，44(9)：58-61.

[39] YAN J，LIU X，QI J，et al. Selective clustering ensemble based on kappa and F-score[EB/OL]. (2022-04-23)[2022-04-25]. https：//arxiv. org/abs/2204. 11062.

[40] VENKATESH M，SATHYALAKSMI S. Memetic swarm clustering with deep belief network model for e-learning recommendation system to improve learning performance[J]. Concurrency and Computation：Practice and Experience，2022，34(18)：e7010. 1-e7010. 21.

[41] STREHL A，GHOSH J. Cluster ensembles：a knowledge reuse framework for combining partitionings[J]. Journal of Machine Learning Research，2002，3：583-617.

[42] VINH N X，EPPS J R，BAILEY J. Information theoretic measures for clusterings comparison：variants，properties，normalization and correction for chance[J]. The Journal of Machine Learning Research，2010，11：2837-2854.

NMF Weighted Ensemble Clustering Algorithm Based on Hierarchical Preprocessing

LI Xiangli[1,2], BI sheng[1,3], WANG Peiyuan[1]

(1. School of Mathematics & Computing Science, Guilin University of Electronic Technology;
2. Guangxi Colleges and Universities Key Laboratory of Data Analysis and Computation;
3. Center for Applied Mathematics of Guangxi (GUET), Guilin Guangxi 541004, China)

Abstract: Image clustering is a hot research topic at present, and nonnegative matrix factorization (NMF) has been widely used in the field of image clustering. However, a single NMF clustering algorithm can't be applied to all datasets, and the NMF algorithm directly processes the original space of the data, which has poor noise resistance. Ensemble clustering can solve the above problems. Ensemble clustering combines several basic clustering results into a consistent result, which not only improves the quality of clustering, but also enhances the robustness of the algorithm. Therefore, a hierarchical preprocessing NMF weighted integrated clustering algorithm is presented. The algorithm introduces the idea of hierarchical division, ensemble clustering and bipartite graph into the NMF algorithm. In the preprocessing stage, the number of clusters is obtained by hierarchical division. The co-association matrix is then obtained by local weighting. Finally, the final clustering result is obtained by partitioning using the consistency function based on the bipartite graph. The algorithm is tested on five datasets to verify the effectiveness of the algorithm over traditional algorithms and other ensemble algorithms.

Keywords: image clustering; ensemble clustering; nonnegative matrix factorization

(责任编辑 黄 颖)